Indirect Cross-validation for Density Estimation

Olga Y. Savchuk, Jeffrey D. Hart, Simon J. Sheather

Abstract

A new method of bandwidth selection for kernel density estimators is proposed. The method, termed *indirect cross-validation*, or ICV, makes use of so-called *selection* kernels. Least squares cross-validation (LSCV) is used to select the bandwidth of a selection-kernel estimator, and this bandwidth is appropriately rescaled for use in a Gaussian kernel estimator. The proposed selection kernels are linear combinations of two Gaussian kernels, and need not be unimodal or positive. Theory is developed showing that the relative error of ICV bandwidths can converge to 0 at a rate of $n^{-1/4}$, which is substantially better than the $n^{-1/10}$ rate of LSCV. Interestingly, the selection kernels that are best for purposes of bandwidth selection are very poor if used to actually estimate the density function. This property appears to be part of the larger and well-documented paradox to the effect that "the harder the estimation problem, the better cross-validation performs." The ICV method uniformly outperforms LSCV in a simulation study, a real data example, and a simulated example in which bandwidths are chosen locally.

KEY WORDS: Kernel density estimation; Bandwidth selection; Cross-validation; Local cross-validation.

1 Introduction

Let $X_1, \ldots, X_n$ be a random sample from an unknown density f. A kernel density estimator of $f(x)$ is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \tag{1}$$

where $h > 0$ is a smoothing parameter, also known as the bandwidth, and K is the kernel, which is generally chosen to be a unimodal probability density function that is symmetric about zero and has finite variance. A popular choice for K is the Gaussian kernel: $\phi(u) = (2\pi)^{-1/2} \exp(-u^2/2)$. To distinguish between estimators with different kernels, we shall refer to estimator (1) with given kernel K as a *K-kernel estimator*. Choosing an appropriate bandwidth is vital for the good performance of a kernel estimate. This paper is concerned with a new method of data-driven bandwidth selection that we call *indirect cross-validation* (ICV).

Many data-driven methods of bandwidth selection have been proposed. The two most widely used are least squares cross-validation, proposed independently by Rudemo (1982) and Bowman (1984), and the Sheather and Jones (1991) plug-in method. Plug-in produces more stable bandwidths than does cross-validation, and hence is the currently more popular method. Nonetheless, an argument can be made for cross-validation since it requires fewer assumptions than plug-in and works well when the density is difficult to estimate; see Loader (1999). A survey of bandwidth selection methods is given by Jones, Marron, and Sheather (1996).

A number of modifications of LSCV has been proposed in an attempt to improve its performance. These include the biased cross-validation method of Scott and Terrell (1987), a method of Chiu (1991a), the trimmed cross-validation of Feluch and Koronacki (1992), the modified cross-validation of Stute (1992), and the method of Ahmad and Ran (2004) based on kernel contrasts. The ICV method is similar in spirit to one-sided cross-

validation (OSCV), which is another modification of cross-validation proposed in the regression context by Hart and Yi (1998). As in OSCV, ICV initially chooses the bandwidth of an L-kernel estimator using least squares cross-validation. Multiplying the bandwidth chosen at this initial stage by a known constant results in a bandwidth, call it $\hat{h}_{ICV}$, that is appropriate for use in a Gaussian kernel estimator.

A popular means of judging a kernel estimator is the mean integrated squared error, i.e., $MISE(h) = E\left[ISE(h)\right]$, where

$$ISE(h) = \int_{-\infty}^{\infty} \left(\hat{f}_h(x) - f(x)\right)^2 \, dx.$$

Letting h_0 be the bandwidth that minimizes $MISE(h)$ when the kernel is Gaussian, we will show that the mean squared error of $\hat{h}_{ICV}$ as an estimator of h_0 converges to 0 at a faster rate than that of the ordinary LSCV bandwidth. We also describe an unexpected bonus associated with ICV, namely that, unlike LSCV, it is robust to rounded data. A fairly extensive simulation study and two data analyses confirm that ICV performs better than ordinary cross-validation in finite samples.

2 Description of indirect cross-validation

We begin with some notation and definitions that will be used subsequently. For an arbitrary function g, define

$$R(g) = \int g(u)^2 \, du, \quad \mu_{jg} = \int u^j g(u) \, du.$$

The LSCV criterion is given by

$$LSCV(h) = R(\hat{f}_h) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{h,-i}(X_i),$$

where, for $i = 1, \ldots, n$, $\hat{f}_{h,-i}$ denotes a kernel estimator using all the original observations except for X_i. When $\hat{f}_h$ uses kernel K, $LSCV$ can be written as

$$\begin{aligned} LSCV(h) &= \frac{1}{nh}R(K) + \frac{1}{n^2 h}\sum_{i \neq j}\int K(t)K\Big(t + \frac{X_i - X_j}{h}\Big)\,dt \\ &\quad - \frac{2}{n(n-1)h}\sum_{i \neq j} K\Big(\frac{X_i - X_j}{h}\Big). \end{aligned} \tag{2}$$

It is well known that $LSCV(h)$ is an unbiased estimator of $MISE(h) - \int f^2(x)\,dx$, and hence the minimizer of $LSCV(h)$ with respect to h is denoted $\hat{h}_{UCV}$.

2.1 The basic method

Our aim is to choose the bandwidth of a *second order* kernel estimator. A second order kernel integrates to 1, has first moment 0, and finite, nonzero second moment. In principle our method can be used to choose the bandwidth of any second order kernel estimator, but in this article we restrict attention to $K \equiv \phi$, the Gaussian kernel. It is well known that a ϕ-kernel estimator has asymptotic mean integrated squared error (MISE) within 5% of the minimum among all positive, second order kernel estimators.

Indirect cross-validation may be described as follows:

- Select the bandwidth of an L-kernel estimator using least squares cross-validation, and call this bandwidth $\hat{b}_{UCV}$. The kernel L is a second order kernel that is a linear combination of two Gaussian kernels, and will be discussed in detail in Section 2.2.

- Assuming that the underlying density f has second derivative which is continuous and square integrable, the bandwidths h_n and b_n that asymptotically minimize the $MISE$ of ϕ- and L-kernel estimators, respectively, are related

as follows:

$$h_n = \left(\frac{R(\phi)\mu_{2L}^2}{R(L)\mu_{2\phi}^2} \right)^{1/5} b_n \equiv C b_n. \tag{3}$$

- Define the indirect cross-validation bandwidth by $\hat{h}_{ICV} = C\hat{b}_{UCV}$. Importantly, the constant C depends on no unknown parameters. Expression (3) and existing cross-validation theory suggest that $\hat{h}_{ICV}/h_0$ will at least converge to 1 in probability, where h_0 is the minimizer of $MISE$ for the ϕ-kernel estimator.

Henceforth, we let $\hat{h}_{UCV}$ denote the bandwidth that minimizes $LSCV(h)$ with $K \equiv \phi$. Theory of Hall and Marron (1987) and Scott and Terrell (1987) shows that the relative error $(\hat{h}_{UCV} - h_0)/h_0$ converges to 0 at the rather disappointing rate of $n^{-1/10}$. In contrast, we will show that $(\hat{h}_{ICV} - h_0)/h_0$ can converge to 0 at the rate $n^{-1/4}$. Kernels L that are sufficient for this result are discussed next.

2.2 Selection kernels

We consider the family of kernels $\mathcal{L} = \{L(\,\cdot\,;\alpha,\sigma) : \alpha \geq 0, \sigma > 0\}$, where, for all u,

$$L(u;\alpha,\sigma) = (1+\alpha)\phi(u) - \frac{\alpha}{\sigma}\phi\left(\frac{u}{\sigma}\right). \tag{4}$$

Note that the Gaussian kernel is a special case of (4) when $\alpha = 0$ or $\sigma = 1$. Each member of $\mathcal{L}$ is symmetric about 0 and such that $\mu_{2L} = \int u^2 L(u)\,du = 1 + \alpha - \alpha\sigma^2$. It follows that kernels in $\mathcal{L}$ are second order, with the exception of those for which $\sigma = \sqrt{(1+\alpha)/\alpha}$.

The family $\mathcal{L}$ can be partitioned into three families: $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$. The first of these is $\mathcal{L}_1 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \sigma < \frac{\alpha}{1+\alpha}\right\}$. Each kernel in $\mathcal{L}_1$ has a negative dip centered at $x = 0$. For α fixed, the smaller σ is, the more extreme the dip; and for

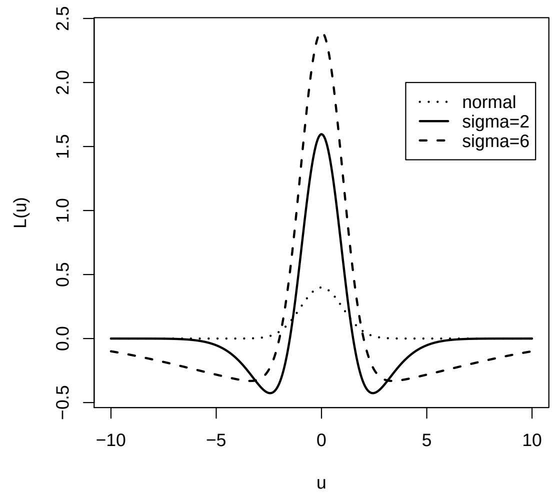

Figure 1: Selection kernels in $\mathcal{L}_3$. The dotted curve corresponds to the Gaussian kernel, and each of the other kernels has $\alpha = 6$.

fixed σ, the larger α is, the more extreme the dip. The kernels in $\mathcal{L}_1$ are ones that "cut-out-the-middle."

The second family is $\mathcal{L}_2 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \frac{\alpha}{1+\alpha} \leq \sigma \leq 1\right\}$. Kernels in $\mathcal{L}_2$ are densities which can be unimodal or bimodal. Note that the Gaussian kernel is a member of this family. The third sub-family is $\mathcal{L}_3 = \left\{L(\cdot;\alpha,\sigma) : \alpha > 0, \sigma > 1\right\}$, each member of which has negative tails. Examples of kernels in $\mathcal{L}_3$ are shown in Figure 1.

Kernels in $\mathcal{L}_1$ and $\mathcal{L}_3$ are not of the type usually used for estimating f. Nonetheless, a worthwhile question is "why not use L for both cross-validation *and* estimation of f?" One could then bypass the step of rescaling $\hat{b}_{UCV}$ and simply estimate f by an L-kernel estimator with bandwidth $\hat{b}_{UCV}$. The ironic answer to this question is that the kernels in $\mathcal{L}$ that are best for cross-validation purposes are very inefficient for estimating f. Indeed, it turns out that an L-kernel estimator based on a sequence

of ICV-optimal kernels has $MISE$ that does not converge to 0 faster than $n^{-1/2}$. In contrast, the $MISE$ of the best ϕ-kernel estimator tends to 0 like $n^{-4/5}$. These facts fit with other cross-validation paradoxes, which include the fact that LSCV outperforms other methods when the density is highly structured, Loader (1999), the improved performance of cross-validation in multivariate density estimation, Sain, Baggerly, and Scott (1994), and its improvement when the true density is not smooth, van Es (1992). One could paraphrase these phenomena as follows: "The more difficult the function is to estimate, the better cross-validation seems to perform." In our work, we have in essence made the function more difficult to estimate by using an inefficient kernel L. More details on the $MISE$ of L-kernel estimators may be found in Savchuk (2009).

3 Large sample theory

The theory presented in this section provides the underpinning for our methodology. We first state a theorem on the asymptotic distribution of $\hat{h}_{ICV}$, and then derive asymptotically optimal choices for the parameters α and σ of the selection kernel.

3.1 Asymptotic mean squared error of the ICV bandwidth

Classical theory of Hall and Marron (1987) and Scott and Terrell (1987) entails that the bias of an LSCV bandwidth is asymptotically negligible in comparison to its standard deviation. We will show that the variance of an ICV bandwidth can converge to 0 at a faster rate than that of an LSCV bandwidth. This comes at the expense of a squared bias that is *not* negligible. However, we will show how to select α and σ (the parameters of the selection kernel) so that the variance and squared bias are balanced and the resulting mean squared error tends to 0 at a faster rate

than does that of the LSCV bandwidth. The optimal rate of convergence of the relative error $(\hat{h}_{ICV} - h_0)/h_0$ is $n^{-1/4}$, a substantial improvement over the infamous $n^{-1/10}$ rate for LSCV.

Before stating our main result concerning the asymptotic distribution of $\hat{h}_{ICV}$, we define some notation:

$$\gamma(u) = \int L(w)L(w+u)\,du - 2L(u), \quad \rho(u) = u\gamma'(u),$$

$$T_n(b) = \sum\sum\nolimits_{1 \le i < j \le n} \left[\gamma\left(\frac{X_i - X_j}{b}\right) + \rho\left(\frac{X_i - X_j}{b}\right) \right],$$

$$T_n^{(j)}(b) = \frac{\partial^j T_n(b)}{\partial b^j}, \quad j = 1, 2,$$

$$A_\alpha = \frac{3}{\sqrt{2\pi}}(1+\alpha)^2 \left[\frac{1}{8}(1+\alpha)^2 - \frac{8}{9\sqrt{3}}(1+\alpha) + \frac{1}{\sqrt{2}} \right],$$

$$C_\alpha = \frac{\sqrt{2A_\alpha}(2\sqrt{\pi})^{9/10}}{5(1+\alpha)^{9/5}\alpha^{1/5}} \quad \text{and} \quad D_\alpha = \frac{3}{20}\left(\frac{(1+\alpha)^2}{2\alpha^2\sqrt{\pi}} \right)^{2/5}.$$

Note that to simplify notation, we have suppressed the fact that L, γ and ρ depend on the parameters α and σ. An outline of the proof of the following theorem is given in the Appendix.

Theorem. *Assume that f and its first five derivatives are continuous and bounded and that $f^{(6)}$ exists and is Lipschitz continuous. Suppose also that*

$$(\hat{b}_{UCV} - b_0)\frac{T_n^{(2)}(\tilde{b})}{T_n^{(1)}(b_0)} = o_p(1) \tag{5}$$

for any sequence of random variables $\tilde{b}$ such that $|\tilde{b} - b_0| \le |\hat{b}_{UCV} - b_0|$, a.s. Then, if $\sigma = o(n)$ and α is fixed,

$$\frac{\hat{h}_{ICV} - h_0}{h_0} = Z_n S_n + B_n + o_p(S_n + B_n),$$

as $n \to \infty$ and $\sigma \to \infty$, where Z_n converges in distribution to a standard normal random variable,

$$S_n = \left(\frac{1}{\sigma^{2/5} n^{1/10}} \right) \frac{R(f)^{1/2}}{R(f'')^{1/10}} C_\alpha, \tag{6}$$

and

$$B_n = \left(\frac{\sigma}{n}\right)^{2/5} \frac{R(f''')}{R(f'')^{7/5}} D_\alpha. \tag{7}$$

Remarks

R1. Assumption (5) is only slightly stronger than assuming that $\hat{b}_{UCV}/b_0$ converges in probability to 1. To avoid making our paper overly technical we have chosen not to investigate sufficient conditions for (5). However, this can be done using techniques as in Hall (1983) and Hall and Marron (1987).

R2. Theorem 4.1 of Scott and Terrell (1987) on asymptotic normality of LSCV bandwidths is not immediately applicable to our setting for at least three reasons: the kernel L is not positive, it does not have compact support, and, most importantly, it changes with n via the parameter σ.

R3. The assumption of six derivatives for f is required for a precise quantification of the asymptotic bias of $\hat{h}_{ICV}$. Our proof of asymptotic normality of $\hat{b}_{UCV}$ only requires that f be four times differentiable, which coincides with the conditions of Theorem 4.1 in Scott and Terrell (1987).

R4. The asymptotic bias B_n is positive, implying that the ICV bandwidth tends to be larger than the optimal bandwidth. This is consistent with our experience in numerous simulations.

In the next section we apply the results of our theorem to determine asymptotically optimal choices for α and σ.

3.2 Minimizing asymptotic mean squared error

The limiting distribution of $(\hat{h}_{ICV} - h_0)/h_0$ has second moment $S_n^2 + B_n^2$, where S_n and B_n are defined by (6) and (7). Minimizing this expression with respect to σ yields the following asymptotically optimal choice for σ:

$$\sigma_{n,opt} = n^{3/8} \left(\frac{C_\alpha}{D_\alpha}\right)^{5/4} \left[\frac{R(f)R(f'')^{13/5}}{R(f''')^2}\right]^{5/8}. \tag{8}$$

The corresponding asymptotically optimal mean squared error is

$$MSE_{n,opt} = n^{-1/2} C_\alpha D_\alpha \left[\frac{R(f''')R(f)^{1/2}}{R(f'')^{3/2}}\right], \tag{9}$$

which confirms our previous claim that the relative error of $\hat{h}_{ICV}$ converges to 0 at the rate $n^{-1/4}$. The corresponding rates for LSCV and the Sheather-Jones plug-in rule are $n^{-1/10}$ and $n^{-5/14}$, respectively.

Because α is not confounded with f in $MSE_{n,opt}$, we may determine a single optimal value of α that is independent of f. The function $C_\alpha D_\alpha$ of α is minimized at $\alpha_0 = 2.4233$. Furthermore, small choices of α lead to an arbitrarily large increase in mean squared error, while the MSE at $\alpha = \infty$ is only about 1.33 times that at the minimum.

Our theory to this point applies to kernels in $\mathcal{L}_3$, i.e., kernels with negative tails. Savchuk (2009) has developed similar theory for the case where $\sigma \to 0$, which corresponds to $L \in \mathcal{L}_1$, i.e., kernels that apply negative weights to the smallest spacings in the LSCV criterion. Interestingly, the same optimal rate of $n^{-1/4}$ results from letting $\sigma \to 0$. However, when the optimal values of (α, σ) are used in the respective cases ($\sigma \to 0$ and $\sigma \to \infty$), the limiting ratio of optimum mean squared errors is 0.752, with $\sigma \to \infty$ yielding the smaller error. Our simulation studies confirm that using L with large σ does lead to more accurate estimation of the optimal bandwidth.

4 Practical choice of α and σ

In order to have an idea of how good choices of α and σ vary with n and f, we determined the minimizers of the asymptotic mean squared error of $\hat{h}_{ICV}$ for various sample sizes and densities. In doing so, we considered a single expression for the asymptotic mean squared error that is valid for either large or small values of σ. Furthermore, we use a slightly enhanced version of the asymptotic bias of $\hat{h}_{ICV}$. The first order bias of $\hat{h}_{ICV}$ is $Cb_0 - h_0$, or $C(b_0 - b_n) + (h_n - h_0)$, where

$$b_n = \left(\frac{R(L)}{\mu_{2L}^2 R(f'')}\right)^{1/5} n^{-1/5} \quad \text{and} \quad h_n = \left(\frac{R(\phi)}{\mu_{2\phi}^2 R(f'')}\right)^{1/5} n^{-1/5}. \tag{10}$$

Now, the term $h_n - h_0$ is of smaller order asymptotically than $C(b_0 - b_n)$ and hence was deleted in the theory of Section 3. Here we retain $h_n - h_0$, and hence the α that minimizes the mean squared error depends on both n and f.

We considered the following five normal mixtures defined in the article by Marron and Wand (1992):

Gaussian density:	$N(0,1)$
Skewed unimodal density:	$\frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$
Bimodal density:	$\frac{1}{2}N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right)$
Separated bimodal density:	$\frac{1}{2}N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
Skewed bimodal density:	$\frac{3}{4}N(0,1) + \frac{1}{4}N\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$.

These choices for f provide a fairly representative range of density shapes. It is worth noting that the asymptotically optimal σ (expression (8)) is free of location and scale. We may thus choose a single representative of a location-scale family when investigating the effect of f. The following remarks summarize our findings about α and σ.

- For each n, the optimal value of σ (α) is larger (smaller) for the unimodal densities than for the bimodal ones.

- All of the MSE-optimal α and σ correspond to kernels from $\mathcal{L}_3$, the family of negative-tailed kernels.

- For each density, the optimal α decreases monotonically with n. Recall from Section 3.2 that the asymptotically optimal α is 2.42. For each unimodal density, the optimal α is within 13.5% of 2.42 at $n = 1000$, and for each bimodal density is within 18% of 2.42 when n is 20,000.

In practice it would be desirable to have choices of α and σ that would adapt to the n and f at hand. However, attempting to estimate optimal values of α and σ is potentially as difficult as the bandwidth selection problem itself. We have built a practical purpose model for α and σ by using polynomial regression. The independent variable was $\log_{10}(n)$ and the dependent variables were the MSE-optimal values of $\log_{10}(\alpha)$ and $\log_{10}(\sigma)$ for the five densities defined above. Using a sixth degree polynomial for α and a quadratic for σ, we arrived at the following models for α and σ:

$$\begin{aligned} \alpha_{mod} &= 10^{3.390 - 1.093 \log 10(n) + 0.025 \log 10(n)^3 - 0.00004 \log 10(n)^6} \\ \sigma_{mod} &= 10^{-0.58 + 0.386 \log 10(n) - 0.012 \log 10(n)^2}, \quad 100 \le n \le 500000. \end{aligned} \tag{11}$$

To the extent that unimodal densities are more prevalent than multimodal densities in practice, these model values are biased towards bimodal cases. Our extensive experience shows that the penalty for using good bimodal choices for α and σ when in fact the density is unimodal, is an increase in the upward bias of $\hat{h}_{ICV}$. Our implementation of ICV, however, guards against oversmoothing by using an objective upper bound on the bandwidth, as we explain in detail in Section 7. We thus feel confident in recommending model (11) for choosing α and σ in practice, at least until a better method is proposed. Indeed, this model is what we used to choose α and σ in the simulation study reported upon in Section 7.

5 Robustness of ICV to data rounding

Silverman (1986, p.52) showed that if the data are rounded to such an extent that the number of pairs $i < j$ for which $X_i = X_j$ is above a threshold, then $LSCV(h)$ approaches $-\infty$ as h approaches zero. This threshold is $0.27n$ for the Gaussian kernel. Chiu (1991b) showed that for data with ties, the behavior of $LSCV(h)$ as $h \to 0$ is determined by the balance between $R(K)$ and $2K(0)$. In particular, $\lim_{h\to 0} LSCV(h)$ is $-\infty$ and ∞ when $R(K) < 2K(0)$ and $R(K) > 2K(0)$, respectively. The former condition holds necessarily if K is nonnegative and has its maximum at 0. This means that all the traditional kernels have the problem of choosing $h = 0$ when the data are rounded.

Recall that selection kernels (4) are not restricted to be nonnegative. It turns out that there exist α and σ such that $R(L) > 2L(0)$ will hold. We say that selection kernels satisfying this condition are robust to rounding. It can be verified that the negative-tailed selection kernels with $\sigma > 1$ are robust to rounding when

$$\alpha > \frac{-a_\sigma + \sqrt{a_\sigma + (2 - 1/\sqrt{2})b_\sigma}}{b_\sigma}, \tag{12}$$

where $a_\sigma = \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{1+\sigma^2}} - 1 + \frac{1}{\sigma}\right)$ and $b_\sigma = \left(\frac{1}{\sqrt{2}} - \frac{2}{\sqrt{1+\sigma^2}} + \frac{1}{\sigma\sqrt{2}}\right)$. It turns out that all the selection kernels corresponding to model (11) are robust to rounding. Figure 2 shows the region (12) and also the curve defined by model (11) for $100 \le n \le 500000$. Interestingly, the boundary separating robust from nonrobust kernels almost coincides with the (α, σ) pairs defined by that model.

6 Local ICV

A local version of cross-validation for density estimation was proposed and analyzed independently by Hall and Schucany (1989) and Mielniczuk, Sarda, and Vieu (1989).

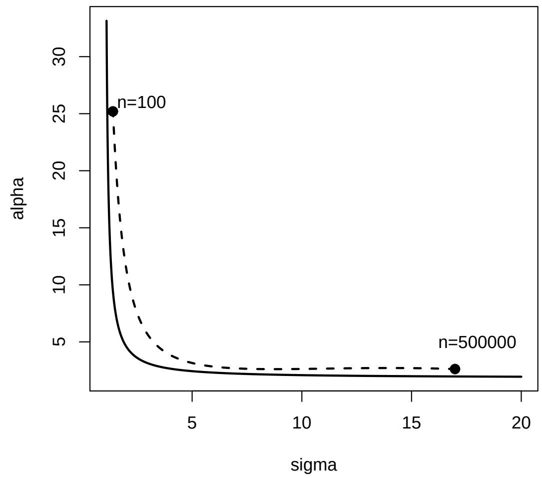

Figure 2: Selection kernels robust to rounding have α and σ above the solid curve. Dashed curve corresponds to the model-based selection kernels.

A local method allows the bandwidth to vary with x, which is desirable when the smoothness of the underlying density varies sufficiently with x. Fan, Hall, Martin, and Patil (1996) proposed a different method of local smoothing that is a hybrid of plug-in and cross-validation methods. Here we propose that ICV be performed locally. The method parallels that of Hall and Schucany (1989) and Mielniczuk, Sarda, and Vieu (1989), with the main difference being that each local bandwidth is chosen by ICV rather than LSCV. We suggest using the *smallest* local minimizer of the ICV curve, since ICV does not have LSCV's tendency to undersmooth.

Let $\hat{f}_b$ be a kernel estimate that employs a kernel in the class $\mathcal{L}$, and define, at the point x, a local ICV curve by

$$ICV(x,b) = \frac{1}{w}\int_{-\infty}^{\infty} \phi\left(\frac{x-u}{w}\right) \hat{f}_b^2(u)\,du - \frac{2}{nw}\sum_{i=1}^{n} \phi\left(\frac{x-X_i}{w}\right) \hat{f}_{b,-i}(X_i), \quad b > 0.$$

The quantity w determines the degree to which the cross-validation is local, with a very large choice of w corresponding to global ICV. Let $\hat{b}(x)$ be the minimizer of

$ICV(x, b)$ with respect to b. Then the bandwidth of a Gaussian kernel estimator at the point x is taken to be $\hat{h}(x) = C\hat{b}(x)$. The constant C is defined by (3), and choice of α and σ in the selection kernel will be discussed in Section 8.

Local LSCV can be criticized on the grounds that, at any x, it promises to be even more unstable than global LSCV since it (effectively) uses only a fraction of the n observations. Because of its much greater stability, ICV seems to be a much more feasible method of local bandwidth selection than does LSCV. We provide evidence of this stability by example in Section 8.

7 Simulation study

The primary goal of our simulation study is to compare ICV with ordinary LSCV. However, we will also include the Sheather-Jones plug-in method in the study. We considered the four sample sizes $n = 100$, 250, 500 and 5000, and sampled from each of the five densities listed in Section 4. For each combination of density and sample size, 1000 replications were performed. Here we give only a synopsis of our results. The reader is referred to Savchuk, Hart, and Sheather (2008) for a much more detailed account of what we observed.

Let $\hat{h}_0$ denote the minimizer of $ISE(h)$ for a Gaussian kernel estimator. For each replication, we computed $\hat{h}_0$, $\hat{h}^*_{ICV}$, $\hat{h}_{UCV}$ and $\hat{h}_{SJPI}$. The definition of $\hat{h}^*_{ICV}$ is $\min(\hat{h}_{ICV}, \hat{h}_{OS})$, where $\hat{h}_{OS}$ is the oversmoothed bandwidth of Terrell (1990). Since $\hat{h}_{ICV}$ tends to be biased upwards, this is a convenient means of limiting the bias. In all cases the parameters α and σ in the selection kernel L were chosen according to model (11). For any random variable Y defined in each replication of our simulation, we denote the average of Y over all replications (with n and f fixed) by $\widehat{E}(Y)$. Our main conclusions may be summarized as follows.

- The ratio $\widehat{E}(\hat{h}^*_{ICV} - \widehat{E}\hat{h}_0)^2/\widehat{E}(\hat{h}_{UCV} - \widehat{E}\hat{h}_0)^2$ ranged between 0.04 and 0.70 in the sixteen settings excluding the skewed bimodal density. For the skewed bimodal, the ratio was 0.84, 1.27, 1.09, and 0.40 at the respective sample sizes 100, 250, 500 and 5000. The fact that this ratio was larger than 1 in two cases was a result of ICV's bias, since the sample standard deviation of the ICV bandwidth was smaller than that for the LSCV bandwidth in all twenty settings.

- The ratio $\widehat{E}\big(ISE(\hat{h}^*_{ICV})/ISE(\hat{h}_0)\big)/\widehat{E}\big(ISE(\hat{h}_{UCV})/ISE(\hat{h}_0)\big)$ was smaller than 1 for every combination of density and sample size. For the two "large bias" cases mentioned in the previous remark the ratio was 0.92.

- The ratio $\widehat{E}\big(ISE(\hat{h}^*_{ICV})/ISE(\hat{h}_0)\big)/\widehat{E}\big(ISE(\hat{h}_{SJPI})/ISE(\hat{h}_0)\big)$ was smaller than 1 in six of the twenty cases considered. Among the other fourteen cases, the ratio was between 1.00 and 1.15, exceeding 1.07 just twice.

- Despite the fact that the LSCV bandwidth is asymptotically normally distributed (see Hall and Marron (1987)), its distribution in finite samples tends to be skewed to the left. In contrast, our simulations show that the ICV bandwidth distribution is nearly symmetric.

8 Examples

In this Section we illustrate the use of ICV with two examples, one involving credit scores from Fannie Mae and the other simulated data. The first example is provided to compare the ICV, LSCV, and Sheather-Jones plug-in methods for choosing a global bandwidth. The second example illustrates the benefit of applying ICV locally.

8.1 Mortgage defaulters

In this example we analyze the credit scores of Fannie Mae clients who defaulted on their loans. The mortgages considered were purchased in "bulk" lots by Fannie Mae from primary banking institutions. The data set was taken from the website `http://www.dataminingbook.com` associated with Shmueli, Patel, and Bruce (2006).

In Figure 3 we have plotted an unsmoothed frequency histogram and the LSCV, ICV and Sheather-Jones plug-in density estimates for the credit scores. The class interval size in the unsmoothed histogram was chosen to be 1, which is equal to the accuracy to which the data have been reported. It turns out that the LSCV curve tends to $-\infty$ when $h \to 0$, but has a local minimum at about 2.84. Using $h = 2.84$ results in a severely undersmoothed estimate. Both the Sheather-Jones plug-in and ICV density estimates show a single mode around 675 and look similar, with the ICV estimate being somewhat smoother. Interestingly, a high percentage of the defaulters have credit scores less than 620, which many lenders consider the minimum score that qualifies for a loan; see Desmond (2008).

8.2 Local ICV: simulated example

For this example we took five samples of size $n = 1500$ from the kurtotic unimodal density defined in Marron and Wand (1992). First, we note that even the bandwidth that minimizes $ISE(h)$ results in a density estimate that is much too wiggly in the tails. On the other hand, using local versions of either ICV or LSCV resulted in much better density estimates, with local ICV producing in each case a visually better estimate than that produced by local LSCV.

For the local LSCV and ICV methods we considered four values of w ranging from 0.05 to 0.3. A selection kernel with $\alpha = 6$ and $\sigma = 6$ was used in local ICV.

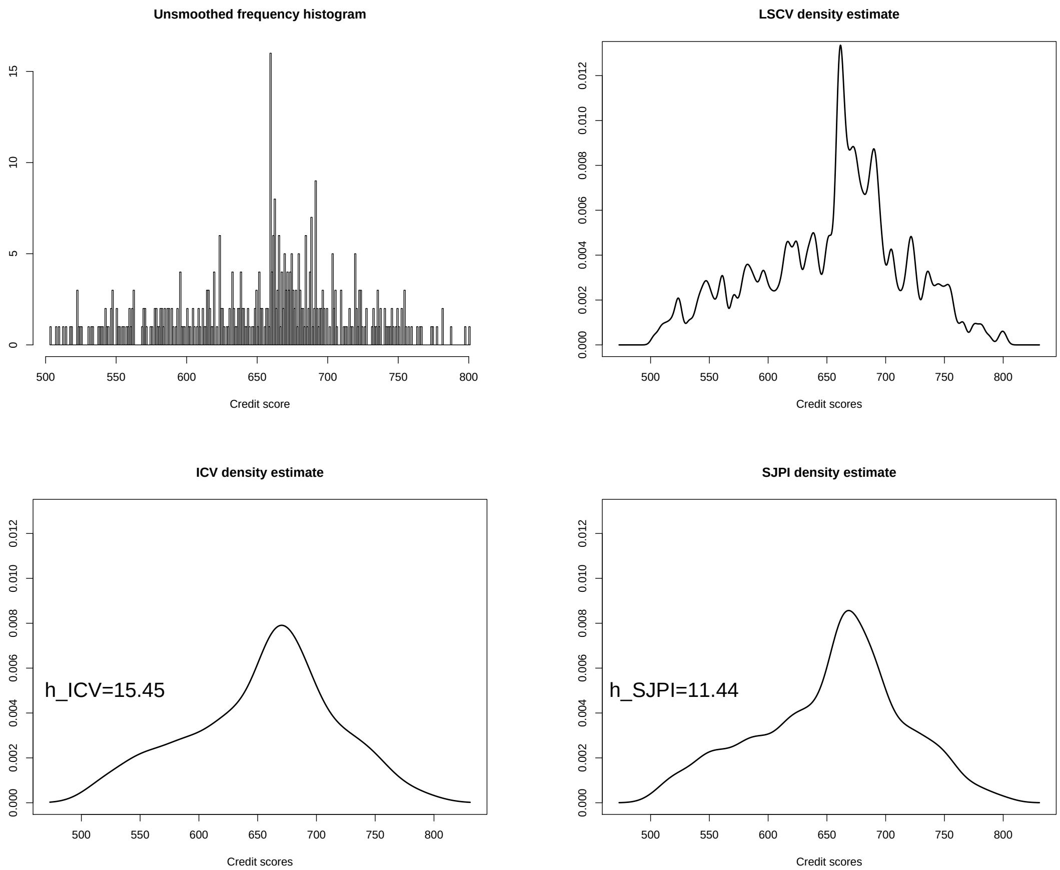

Figure 3: Unsmoothed histogram and kernel density estimates for credit scores.

This (α, σ) choice performs well for global bandwidth selection when the density is unimodal, and hence seems reasonable for local bandwidth selection since locally the density should have relatively few features. For a given w, the local ICV and LSCV bandwidths were found for $x = -3, -2.9, \ldots, 2.9, 3$, and were interpolated at other $x \in [-3, 3]$ using a spline. Average squared error (ASE) was used to measure closeness of a local density estimate $\hat{f}_\ell$ to the true density f:

$$ASE = \frac{1}{61} \sum_{i=1}^{61} (\hat{f}_\ell(x_i) - f(x_i))^2.$$

Figure 4 shows results for one of the five samples. Estimates corresponding to the smallest and the largest values of w are provided. The local ICV method performed

similarly well for all values of w considered, whereas all the local LSCV estimates were very unsmooth, albeit with some improvement in smoothness as w increased.

9 Summary

A widely held view is that kernel choice is not terribly important when it comes to estimation of the underlying curve. In this paper we have shown that kernel choice can have a dramatic effect on the properties of cross-validation. Cross-validating kernel estimates that use Gaussian or other traditional kernels results in highly variable bandwidths, a result that has been well-known since at least 1987. We have shown that certain kernels with low efficiency for estimating f can produce cross-validation bandwidths whose relative error converges to 0 at a faster rate than that of Gaussian-kernel cross-validation bandwidths.

The kernels we have studied have the form $(1+\alpha)\phi(u) - \alpha\phi(u/\sigma)/\sigma$, where ϕ is the standard normal density and α and σ are positive constants. The interesting selection kernels in this class are of two types: unimodal, negative-tailed kernels and "cut-out the middle kernels," i.e., bimodal kernels that go negative between the modes. Both types of kernels yield the rate improvement mentioned in the previous paragraph. However, the best negative-tailed kernels yield bandwidths with smaller asymptotic mean squared error than do the best "cut-out-the-middle" kernels.

A model for choosing the selection kernel parameters has been developed. Use of this model makes our method completely automatic. A simulation study and examples reveal that use of this method leads to improved performance relative to ordinary LSCV.

To date we have considered only selection kernels that are a linear combination of two normal densities. It is entirely possible that another class of kernels would work

even better. In particular, a question of at least theoretical interest is whether or not the convergence rate of $n^{-1/4}$ for the relative bandwidth error can be improved upon.

10 Appendix

Here we outline the proof of our theorem in Section 3. A much more detailed proof is available from the authors.

We start by writing

$$\begin{aligned} T_n(b_0) &= T_n(\hat{b}_{UCV}) + (b_0 - \hat{b}_{UCV})T_n^{(1)}(b_0) + \frac{1}{2}(b_0 - \hat{b}_{UCV})^2 T_n^{(2)}(\tilde{b}) \\ &= -nR(L)/2 + (b_0 - \hat{b}_{UCV})T_n^{(1)}(b_0) + \frac{1}{2}(b_0 - \hat{b}_{UCV})^2 T_n^{(2)}(\tilde{b}), \end{aligned}$$

where $\tilde{b}$ is between b_0 and $\hat{b}_{UCV}$, and so

$$(\hat{b}_{UCV} - b_0)\left(1 - (\hat{b}_{UCV} - b_0)\frac{T_n^{(2)}(\tilde{b})}{2T_n^{(1)}(b_0)}\right) = \frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)}.$$

Using condition (5) we may write the last equation as

$$(\hat{b}_{UCV} - b_0) = \frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)} + o_p\left(\frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)}\right). \tag{13}$$

Defining $s_n^2 = \text{Var}(T_n(b_0))$ and $\beta_n = E(T_n(b_0)) + nR(L)/2$, we have

$$\frac{T_n(b_0) + nR(L)/2}{-T_n^{(1)}(b_0)} = \frac{T_n(b_0) - ET_n(b_0)}{s_n} \cdot \frac{s_n}{-T_n^{(1)}(b_0)} + \frac{\beta_n}{-T_n^{(1)}(b_0)}.$$

Using the central limit theorem of Hall (1984), it can be verified that

$$Z_n \equiv \frac{T_n(b_0) - ET_n(b_0)}{s_n} \xrightarrow{\mathcal{D}} N(0,1).$$

Computation of the first two moments of $T_n^{(1)}(b_0)$ reveals that

$$\frac{-T_n^{(1)}(b_0)}{5R(f'')b_0^4\mu_{2L}^2 n^2/2} \xrightarrow{p} 1,$$

and so

$$\frac{T_n(b_0)+nR(L)/2}{-T_n^{(1)}(b_0)} = Z_n \cdot \frac{2s_n}{5R(f'')b_0^4\mu_{2L}^2n^2} + \frac{2\beta_n}{5R(f'')b_0^4\mu_{2L}^2n^2} + o_p\left(\frac{s_n+\beta_n}{b_0^4\mu_{2L}^2n^2}\right).$$

At this point we need the first two moments of $T_n(b_0)$. A fact that will be used frequently from this point on is that $\mu_{2k,L} = O(\sigma^{2k})$, $k = 1, 2, \ldots$. Using our assumptions on the smoothness of f, Taylor series expansions, symmetry of γ about 0 and $\mu_{2\gamma} = 0$,

$$ET_n(b_0) = -\frac{n^2}{12}b_0^5\mu_{4\gamma}R(f'') + \frac{n^2}{240}b_0^7\mu_{6\gamma}R(f''') + O(n^2b_0^8\sigma^7).$$

Recalling the definition of b_n from (10), we have

$$\begin{aligned}\beta_n &= -\frac{n^2}{12}b_0^5\mu_{4\gamma}R(f'') + \frac{n^2}{240}b_0^7\mu_{6\gamma}R(f''') \\ &\quad + \frac{n^2}{2}b_n^5\mu_{2L}^2R(f'') + O(n^2b_0^8\sigma^7). \qquad (14)\end{aligned}$$

Let $MISE_L(b)$ denote the MISE of an L-kernel estimator with bandwidth b. Then $MISE_L'(b_n) = (b_n - b_0)MISE_L''(b_0) + o\left[(b_n - b_0)MISE_L''(b_0)\right]$, implying that

$$b_n^5 = b_0^5 + 5b_0^4\frac{MISE_L'(b_n)}{MISE_L''(b_0)} + o\left[b_0^4\frac{MISE_L'(b_n)}{MISE_L''(b_0)}\right]. \qquad (15)$$

Using a second order approximation to $MISE_L'(b)$ and a first order approximation to $MISE_L''(b)$, we then have

$$b_n^5 = b_0^5 - b_0^7\frac{\mu_{2L}\mu_{4L}R(f''')}{4\mu_{2L}^2R(f'')} + o(b_0^7\sigma^2).$$

Substitution of this expression for b_n into (14) and using the facts $\mu_{4\gamma} = 6\mu_{2L}^2$, $\mu_{6\gamma} = 30\mu_{2L}\mu_{4L}$ and $b_0\sigma = o(1)$, it follows that $\beta_n = o(n^2b_0^7\sigma^6)$. Later in the proof we will see that this last result implies that the first order bias of $\hat{h}_{ICV}$ is due only to the difference $Cb_0 - h_0$.

Tedious but straightforward calculations show that $s_n^2 \sim n^2b_0R(f)A_\alpha/2$, where A_α is as defined in Section 3.1. It is worth noting that $A_\alpha = R(\rho_\alpha)$, where $\rho_\alpha(u) =$

$u\gamma_\alpha'(u)$ and $\gamma_\alpha(u) = (1+\alpha)^2 \int \phi(u+v)\phi(v)\,dv - 2(1+\alpha)\phi(u)$. One would expect from Theorem 4.1 of Scott and Terrell (1987) that the factor $R(\rho)$ would appear in $\mathrm{Var}(T_n(b_0))$. Indeed it does implicitly, since $R(\rho_\alpha) \sim R(\rho)$ as $\sigma \to \infty$. Our point is that, when $\sigma \to \infty$, the part of L depending on σ is negligible in terms of its effect on $R(\rho)$ and also $R(L)$.

To complete the proof write

$$\begin{aligned}
\frac{\hat{h}_{ICV} - h_0}{h_0} &= \frac{\hat{h}_{ICV} - h_0}{h_n} + o_p\left[\frac{\hat{h}_{ICV} - h_0}{h_n}\right] \\
&= \frac{\hat{b}_{UCV} - b_0}{b_n} + \frac{(Cb_0 - h_0)}{h_n} + o_p\left[\frac{\hat{h}_{ICV} - h_0}{h_n}\right].
\end{aligned}$$

Applying the same approximation of b_0 that led to (15), and the analogous one for h_0, we have

$$\begin{aligned}
\frac{Cb_0 - h_0}{h_n} &= b_n^2 \frac{\mu_{2L}\mu_{4L}R(f''')}{20\mu_{2L}^2 R(f'')} - h_n^2 \frac{\mu_{2\phi}\mu_{4\phi}R(f''')}{20\mu_{2\phi}^2 R(f'')} + o(b_n^2\sigma^2 + h_n^2) \\
&= \frac{R(L)^{2/5}\mu_{2L}\mu_{4L}R(f''')}{20(\mu_{2L}^2)^{7/5}R(f'')^{7/5}}\, n^{-2/5} + o(b_n^2\sigma^2).
\end{aligned}$$

It is easily verified that, as $\sigma \to \infty$, $R(L) \sim (1+\alpha)^2/(2\sqrt{\pi})$, $\mu_{2L} \sim -\alpha\sigma^2$ and $\mu_{4L} \sim -3\alpha\sigma^4$, and hence

$$\frac{Cb_0 - h_0}{h_n} = \left(\frac{\sigma}{n}\right)^{2/5} \frac{R(f''')}{R(f'')^{7/5}} D_\alpha + o\left[\left(\frac{\sigma}{n}\right)^{2/5}\right].$$

The proof is now complete upon combining all the previous results.

11 Acknowledgements

The authors are grateful to David Scott and George Terrell for providing valuable insight about cross-validation, and to three referees and an associate editor, whose comments led to a much improved final version of our paper. The research of Savchuk and Hart was supported in part by NSF Grant DMS-0604801.

References

Ahmad, I. A. and I. S. Ran (2004). Kernel contrasts: a data-based method of choosing smoothing parameters in nonparametric density estimation. *J. Nonparametr. Stat. 16*(5), 671–707.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*(2), 353–360.

Chiu, S.-T. (1991a). Bandwidth selection for kernel density estimation. *Ann. Statist. 19*(4), 1883–1905.

Chiu, S.-T. (1991b). The effect of discretization error on bandwith selection for kernel density estimation. *Biometrika 78*(2), 436–441.

Desmond, M. (2008). Lipstick on a pig. *Forbes*.

Fan, J., P. Hall, M. A. Martin, and P. Patil (1996). On local smoothing of nonparametric curve estimators. *J. Amer. Statist. Assoc. 91*(433), 258–266.

Feluch, W. and J. Koronacki (1992). A note on modified cross-validation in density estimation. *Comput. Statist. Data Anal. 13*(2), 143–151.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist. 11*(4), 1156–1174.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal. 14*(1), 1–16.

Hall, P. and J. S. Marron (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields 74*(4), 567–581.

Hall, P. and W. R. Schucany (1989). A local cross-validation algorithm. *Statist. Probab. Lett. 8*(2), 109–117.

Hart, J. D. and S. Yi (1998). One-sided cross-validation. *J. Amer. Statist. Assoc. 93*(442), 620–631.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc. 91*(433), 401–407.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist. 27*(2), 415–438.

Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *Ann. Statist. 20*(2), 712–736.

Mielniczuk, J., P. Sarda, and P. Vieu (1989). Local data-driven bandwidth choice for density estimation. *J. Statist. Plann. Inference 23*(1), 53–69.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist. 9*(2), 65–78.

Sain, S. R., K. A. Baggerly, and D. W. Scott (1994). Cross-validation of multivariate densities. *J. Amer. Statist. Assoc. 89*(427), 807–817.

Savchuk, O. (2009). *Choosing a kernel for cross-validation*. PhD thesis, Texas A&M University.

Savchuk, O. Y., J. D. Hart, and S. J. Sheather (2008). An empirical study of indirect cross-validation. *Festschrift for Tom Hettmansperger. IMS Lecture Notes-Monograph Series*. Submitted.

Scott, D. W. and G. R. Terrell (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc. 82*(400), 1131–1146.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B 53*(3), 683–690.

Shmueli, G., N. R. Patel, and P. C. Bruce (2006). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. New York: Wiley.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Stute, W. (1992). Modified cross-validation in density estimation. *J. Statist. Plann. Inference 30*(3), 293–305.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc. 85*(410), 470–477.

van Es, B. (1992). Asymptotics for least squares cross-validation bandwidths in nonsmooth cases. *Ann. Statist. 20*(3), 1647–1657.

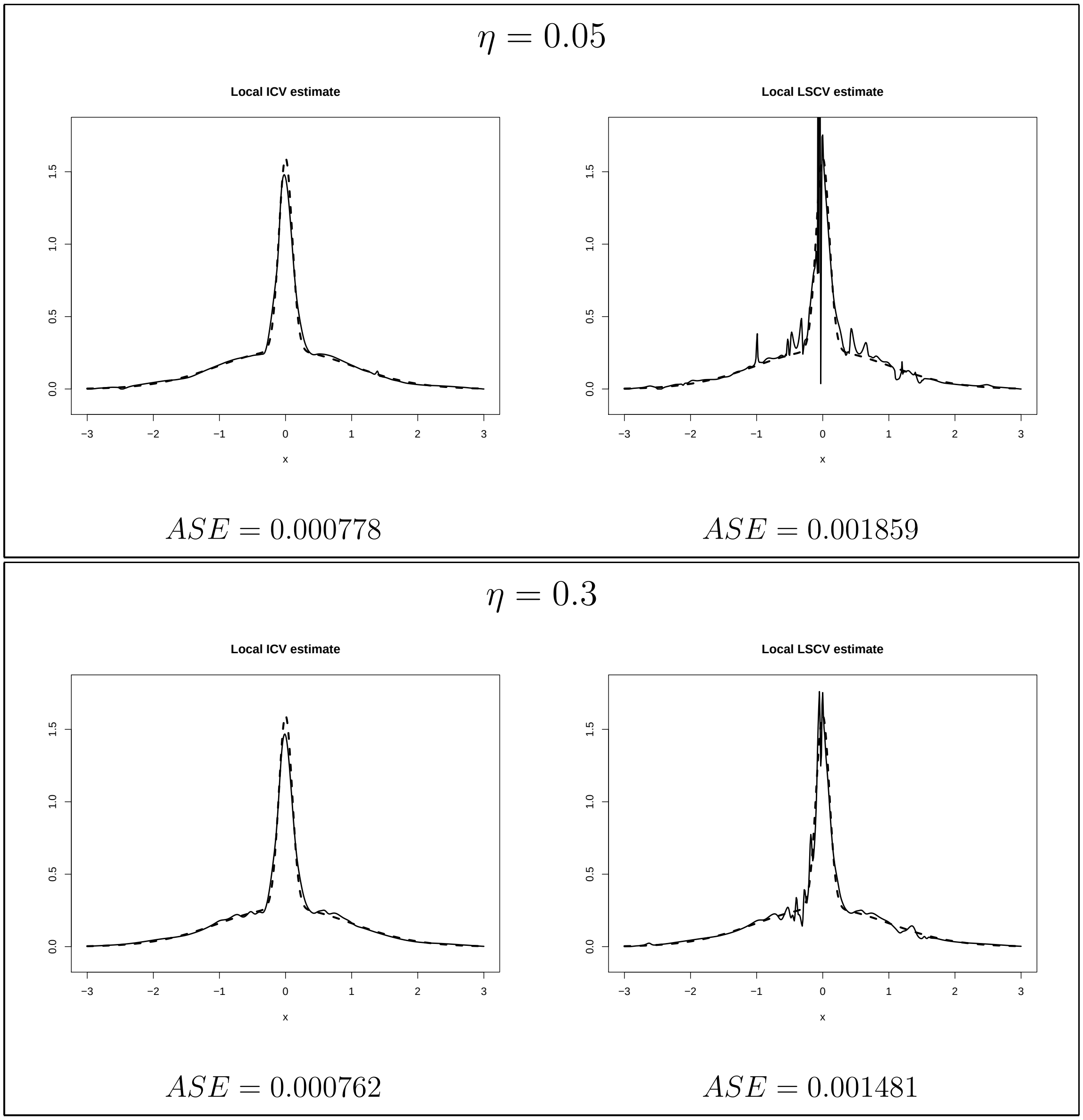

Figure 4: The solid curves correspond to the local LSCV and ICV density estimates, whereas the dashed curves show the kurtotic unimodal density.